Determination of Equilibrium Constants with the OPIUM Computer Program

Mojmír Kývala[a], Přemek Lubal[b], Ivan Lukeš[a]

[a]Department of Inorganic Chemistry, Charles University, Hlavova 2030, 128 43 Praha 2, Czech Republic

[b]Department of Analytical Chemistry, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic

The recently published[1] OPIUM computer program for the analysis of solution equilibria by the generalized least squares method is now compared with several well-known programs for the analysis of potentiometric, spectrophotometric (or enthalpimetric) and NMR chemical shift (or capillary zone electrophoresis effective mobility) data. To increase the objectiveness of the comparison, only "third party" data are treated (formation of polynuclear complexes, ligand competition, dissociation of a weak acid).

The main feature of the OPIUM program (besides extensive analysis of both the estimates and the residuals¶) is a strict separation of the dependence of a response variable y on the calibration and chemical parameters $y = f(\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}, x)$ into two *independent* parts. These are a *calibration function* (the dependence of the response variable y on the calibration parameters $\boldsymbol{\vartheta}$ and the free concentrations $\boldsymbol{c}$ of all components) $y = \phi(\boldsymbol{\vartheta}, \boldsymbol{c})$ and the system of nonlinear equations (mass balances)

$$c_k + \sum_{i=1}^{m} q_{ik}\beta_i \prod_{j=1}^{n} c_j^{q_{ij}} = \frac{\mathrm{c}_k \mathrm{V} + \mathrm{c}'_k x}{\mathrm{V} + x} = \frac{\mathrm{n}_k + \mathrm{c}'_k x}{\mathrm{V} + x} \quad , \qquad k = 1, \ldots, n \,,$$

n is the number of reactants, m is the number of complexes, c_j is the free concentration of the j-th reactant, q_{ij} is the element of the matrix of stoichiometric coefficients $\mathbf{q} \sim (m, n)$, β_i is the stability constant of the i-th complex, c_j (n_j) is the analytical concentration (mass amount) of the j-th reactant in the titrand, V is the volume of the titrand, c'_j is the analytical concentration of the j-th reactant in the titrant and x is the added volume of the titrant (only if a titration is applied). This separation (together with a correct definition of the variance-covariance matrix $\mathbf{V}$ of the random vector of the dependent variable $\mathbf{y}$) is a key to the treatment of many different types of data in any combination.

Formally, any subset of the parameters $\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}$ can be "marked" *unknown* and calculated. Moreover, simple boundary conditions for selected unknown parameters like for instance $a_1 = a_2 + a_3$† or $\mathrm{c}_1 = 0.0736 + 4\mathrm{c}_2 \wedge \mathrm{c}_3 = \mathrm{c}_2$ can be employed to suppress the correlation of the unknown parameters esti-

¶ including their bias and implied parametric and internal curvature of the model

† a_1, a_2, a_3 could be e.g. molar absorption coefficients

mates. The OPIUM program also analyzes data containing sections with *zero* free concentrations of some reactants (necessary, for example, when calculating dissociation constants of a weak acid and stability constants of complexes of the acid with several metals together, i.e. from both types of titrations).

It is proved, that the OPIUM program is robust enough to mostly find the global minimum using automatically generated initial estimates of the unknown stability constants¶ (i.e. without any human assistance). This feature allows us to test many thousands of chemical model hypotheses by the "trial and error" method in several hours† (the program is written in the standard ANSI C language and can be run on virtually any computer). The best model(s) can be, then, verified using the built-in ESI (estimation of stoichiometric indices) approach[2], i.e. a joint calculation of all the unknown stability constants and selected elements $q_{ij} \in \mathsf{R}$ of the matrix of stoichiometric coefficients $\mathbf{q}$.

1 Description of the OPIUM Computer Program

1.1 Chemical model (mass balances)

To be a little more accurate, we have to define the number of reactants with unknown free concentration $r \leq n$ (a lot of spectroscopic data are collected using buffered solutions, i.e. the free concentration instead of the analytical concentration of H^+ is known!). Then the mass balances are§

$$f_k = c_k + \sum_{i=1}^{m} q_{ik} c_{n+i} - \mathrm{C}_k = 0 \quad , \quad k = 1, \ldots, r \ ,$$

$$c_{n+i} = \beta_i \prod_{j=1}^{n} c_j^{q_{ij}} = \beta_i e^{\sum_{j=1}^{n} q_{ij} \ln c_j} = \beta_i 10^{\sum_{j=1}^{n} q_{ij} \log c_j} = 10^{\log \beta_i + \sum_{j=1}^{n} q_{ij} \log c_j} \quad , \quad i = 1, \ldots, m \ .$$

Partial derivatives of the components of the of the function $\mathbf{f}$ with respect to the logarithms of the unknown free concentrations of reactants are

$$\frac{\partial f_k}{\partial \ln c_k} = 1 + \sum_{i=1}^{m} q_{ik}^2 c_{n+i} \quad , \quad k = 1, \ldots, r \ ,$$

$$\frac{\partial f_k}{\partial \ln c_l} = \frac{\partial f_l}{\partial \ln c_k} = \sum_{i=1}^{m} q_{ik} q_{il} c_{n+i} \quad , \quad k \neq l = 1, \ldots, r \ .$$

¶ If initial estimates of the unknown stability constants are not given, the OPIUM program generates them according to the simple formula $\log \beta_i = A + B \sum_{j=1}^{n} q_{ij}$, A and B are empirical constants. If all the other chemical parameters $\mathbf{c}$, $\mathbf{n}$, V, $\mathbf{c}$', $\mathbf{q}$ are *known*, the unknown parameters optimization method used usually ensures rapid convergence to the correct solution. If they are **not**, some other local minima often exist and better initial estimates of $\boldsymbol{\beta}$ might be required to find the "correct" one.

† while being somewhere else and/or doing something more exciting

§ let the $n + m$ components are ordered such that the first r of them are reactants with unknown free (known analytical) concentrations, the next $n - r$ are reactants with known free (unknown analytical) concentrations and the last m are complexes

Partial derivatives of the function $y = f(\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c'}, \mathbf{q}, \boldsymbol{\beta}, x)$ with respect to the *unknown* chemical parameters $\mathbf{p} \equiv \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c'}, \mathbf{q}, \boldsymbol{\beta}$ are¶

$$\frac{\partial y}{\partial p} = \sum_{i=1}^{n+m} \frac{\partial y}{\partial c_i} \frac{\partial c_i}{\partial p} \ .$$

Derivatives $\frac{\partial c_i}{\partial p}$, $i = 1, \ldots, r$, are the solution of the system of linear equations

$$\frac{\partial c_k}{\partial p} + \sum_{l=1}^{n} \frac{\partial c_l}{\partial p} \frac{1}{c_l} \sum_{\substack{i=1 \\ q_{il} \neq 0}}^{m} q_{ik} q_{il} c_{n+i} = A_k \quad , \quad k = 1, \ldots, r \ ,$$

for $p \equiv \mathrm{c}_v, \mathrm{n}_v, \mathrm{V}, \mathrm{c}'_v$

$$A_k = \frac{\partial \mathrm{C}_k}{\partial p} \ \dagger ,$$

for $p \equiv \beta_u$

$$A_k = -q_{uk} \prod_{j=1}^{n} c_j^{q_{uj}} \ ,$$

for $p \equiv q_{uv} \wedge k \neq v$

$$A_k = -q_{uk} \ln c_v \, c_{n+u} \ ,$$

for $p \equiv q_{uk}$

$$A_k = -(1 + q_{uk} \ln c_k) c_{n+u} \ .$$

Derivatives $\frac{\partial c_i}{\partial p}$, $i = r+1, \ldots, n$, are *zero*§, derivatives $\frac{\partial c_i}{\partial p}$, $i = n+1, \ldots, n+m$ (i.e. $\frac{\partial c_{n+i}}{\partial p}$, $i = 1, \ldots, m$), are

$$\frac{\partial c_{n+i}}{\partial p} = B_i + \sum_{\substack{k=1 \\ q_{ik} \neq 0}}^{n} \frac{\partial c_k}{\partial p} \frac{1}{c_k} q_{ik} c_{n+i} \quad , \quad i = 1, \ldots, m \ ,$$

¶ excluding two exceptions (see the description of calibration functions of types 1 and −5)

† it is clear that

$$\mathrm{C}_k = \frac{\mathrm{c}_k \mathrm{V} + \mathrm{c}'_k x}{\mathrm{V} + x} = \frac{\mathrm{n}_k + \mathrm{c}'_k x}{\mathrm{V} + x} \ ,$$

$$\frac{\partial \mathrm{C}_k}{\partial \mathrm{c}_v} = \frac{\partial \mathrm{C}_k}{\partial \mathrm{n}_v} = \frac{\partial \mathrm{C}_k}{\partial \mathrm{c}'_v} = 0 \quad , \quad v \neq k = 1, \ldots, r \ ,$$

$$\frac{\partial \mathrm{C}_k}{\partial \mathrm{c}_k} = \frac{\mathrm{V}}{\mathrm{V} + x} \quad , \quad \frac{\partial \mathrm{C}_k}{\partial \mathrm{n}_k} = \frac{1}{\mathrm{V} + x} \quad , \quad \frac{\partial \mathrm{C}_k}{\partial \mathrm{c}'_k} = \frac{x}{\mathrm{V} + x} \ ,$$

$$\left(\frac{\partial \mathrm{C}_k}{\partial \mathrm{V}} \right)_{\mathrm{dc}_k = 0} = \frac{(\mathrm{c}_k - \mathrm{c}'_k) x}{(\mathrm{V} + x)^2} \quad , \quad \left(\frac{\partial \mathrm{C}_k}{\partial \mathrm{V}} \right)_{\mathrm{dn}_k = \mathrm{dc}_k \mathrm{V} + \mathrm{c}_k d\mathrm{V} = 0} = -\frac{\mathrm{n}_k + \mathrm{c}'_k x}{(\mathrm{V} + x)^2}$$

§ This is **not** true, if the "known" free concentration of the i-th reactant is considered an *unknown* parameter! The OPIUM program provides this option, if an alternative definition of a solution is used (solutions are **not** prepared by a titration).

for $p \equiv \beta_i$

$$B_i = \prod_{j=1}^{n} c_j^{q_{ij}} \ ,$$

for $p \equiv q_{ik}$

$$B_i = \ln c_k \, c_{n+i} \ ,$$

$B_i = 0$ for any other p.

1.2 Measurements (calibration functions)

Overview of the supported calibration functions follows¶.

$\boxed{\textbf{Type } 1}$

$$y = E_0 + S \log c_1 + j_1 c_1 + j_2 \frac{\beta_1}{c_1}$$

$\boldsymbol{\vartheta} \equiv E_0,\ S,\ j_1,\ j_2$

$$\frac{\partial y}{\partial c_1} = \frac{S}{\ln 10 \cdot c_1} + j_1 - j_2 \frac{\beta_1}{c_1^2} \quad , \quad \frac{\partial y}{\partial \beta_1} = \frac{\partial y}{\partial c_1}\frac{\partial c_1}{\partial \beta_1} + \frac{j_2}{c_1}$$

Note: Function of type 1 describes the dependence of the difference y of potentials of a "hydrogen" (glass) electrode and a reference electrode in a cell with non-negligible liquid-junction potential on the free concentration c_1 ($\beta_1 \equiv K$ is the ionic product of water). Thus, H^+ must be the 1st reactant (1st column of the matrix $\mathbf{q}$), OH^- must be the 1st complex (1st row of the matrix $\mathbf{q}$).

$\boxed{\textbf{Type } 2}$

$$y = E_0 + S \log c_1$$

$\boldsymbol{\vartheta} \equiv E_0,\ S$

$\boxed{\textbf{Type } 3}$

$$y = E_0 + S \log c_k + j_1 c_1 + j_2 \frac{\beta_1}{c_1}$$

$\boldsymbol{\vartheta} \equiv E_0,\ S,\ j_1,\ j_2;\ k \in \langle 1, n \rangle$

$\boxed{\textbf{Type } 4}$

$$y = E_0 + S \log c_k$$

$\boldsymbol{\vartheta} \equiv E_0,\ S;\ k \in \langle 1, n \rangle$

$\boxed{\textbf{Type } 5}$

$$y = E_0 + S \log \frac{c_k}{c_l} + j_1 c_1 + j_2 \frac{\beta_1}{c_1}$$

$\boldsymbol{\vartheta} \equiv E_0,\ S,\ j_1,\ j_2;\ k \neq l \in \langle 1, n \rangle$

$\boxed{\textbf{Type } 6}$

$$y = E_0 + S \log \frac{c_k}{c_l}$$

¶ including some "non-trivial" derivatives

$\boldsymbol{\vartheta} \equiv E_0,\ S;\ k \neq l \in \langle 1, n \rangle$

$$\frac{\partial y}{\partial c_k} = \frac{S}{\ln 10 \cdot c_k} \quad , \quad \frac{\partial y}{\partial c_l} = -\frac{S}{\ln 10 \cdot c_l}$$

Note: Functions of types 5 and 6 describe the dependence of the difference y of potentials of a redox (platinum) electrode and a reference electrode on the ratio of free concentrations c_k, c_l of two reactants (i.e. the simplest case of a redox system, e.g. Fe^{3+}/Fe^{2+}, Ce^{4+}/Ce^{3+} etc.).

$\boxed{\textbf{Type } 7}$

$$y = \sum_{i=1}^{n+m} a_i c_i = \mathbf{a}^{\mathrm{T}} \boldsymbol{c}$$

$\boldsymbol{\vartheta} \equiv a_i,\ i = 1, \ldots, n+m$

$$\frac{\partial y}{\partial a_i} = c_i \quad , \quad \frac{\partial y}{\partial c_i} = a_i \quad , \quad i = 1, \ldots, n+m$$

$\boxed{\textbf{Type } 8}$

$$y = d \sum_{i=1}^{n+m} a_i c_i = d\, \mathbf{a}^{\mathrm{T}} \boldsymbol{c}$$

$\boldsymbol{\vartheta} \equiv a_i,\ i = 1, \ldots, n+m;\ d \in \mathsf{R}$

Note: Functions of types 7 and 8 describe the dependence of an extensive *linearly additive* quantity y (absorbance in some "slow exchange" spectrometry (in the case of inseparable peaks), usually UV/VIS spectrophotometry, reaction heat in enthalpimetry, conductivity in conductimetry etc.) on the free concentrations $\boldsymbol{c}$ of all components.

$\boxed{\textbf{Type } 9}$

$$y = \frac{c_i}{c_j}$$

$i \neq j \in \langle 1, n+m \rangle$

$$\frac{\partial y}{\partial c_i} = \frac{1}{c_j} \quad , \quad \frac{\partial y}{\partial c_j} = -\frac{c_i}{c_j^2}$$

$\boxed{\textbf{Type } -1}$

$$y = -\log c_1$$

$\boxed{\textbf{Type } -2}$

$$y = -\log c_k$$

$k \in \langle 1, n \rangle$

$\boxed{\textbf{Type } -3}$

$$y = \alpha c_i$$

$\vartheta \equiv \alpha;\ i \in \langle 1, n+m \rangle$

$\boxed{\textbf{Type } -4}$

$$y = d\alpha c_i$$

$\vartheta \equiv \alpha$; $d \in \mathsf{R}$, $i \in \langle 1, n+m \rangle$

$\boxed{\textbf{Type } -5}$

$$y = \sum_{i=1}^{n+m} a_i p_i c_i / \mathcal{C} = \mathbf{a}^{\mathrm{T}} \mathbf{P} \boldsymbol{c} / \mathcal{C} \ \P\dagger$$

$$\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_{n+m})$$

$\boldsymbol{\vartheta} \equiv a_i$, $i = 1, \ldots, n+m$

$$\frac{\partial y}{\partial a_i} = p_i c_i / \mathcal{C} \quad , \quad \frac{\partial y}{\partial c_i} = p_i \left(a_i \mathcal{C} - \sum_{j=1}^{n+m} a_j p_j c_j \right) / \mathcal{C}^2 = p_i \left(a_i - y \right) / \mathcal{C}$$

$$i = 1, \ldots, n+m$$

$$\frac{\partial y}{\partial q_{uv}} = \sum_{i=1}^{n+m} \frac{\partial y}{\partial c_i} \frac{\partial c_i}{\partial q_{uv}} + p_v c_{n+u} \left(a_{n+u} - y \right) / \mathcal{C} \quad , \quad u = 1, \ldots, m \quad , \quad v = 1, \ldots, n$$

Note: Function of type -5 describes the dependence of an <u>intensive</u> *linearly additive* quantity y (chemical shift or relaxation rate, i.e. reciprocal relaxation time, in "fast exchange" NMR spectroscopy, effective mobility in capillary zone electrophoresis of a labile system etc.) on the free concentrations $\boldsymbol{c}$ of all components.

$\boxed{\textbf{Type } -6}$

$$y = p_i c_i / \mathcal{C} \ \P\dagger$$

$i \in \langle 1, n+m \rangle$

$$\frac{\partial y}{\partial c_i} = p_i (\mathcal{C} - p_i c_i) / \mathcal{C}^2 = p_i (1 - y) / \mathcal{C}$$

$$\frac{\partial y}{\partial c_j} = -p_j p_i c_i / \mathcal{C}^2 = -p_j y / \mathcal{C} \quad , \quad j = 1, \ldots, n+m \ \wedge \ j \neq i$$

$$\frac{\partial y}{\partial q_{uv}} = \sum_{j=1}^{n+m} \frac{\partial y}{\partial c_j} \frac{\partial c_j}{\partial q_{uv}} + A_{uv}$$

¶

$$p_i \in \mathsf{N} \quad , \quad i = 1, \ldots, n \quad , \quad p_{n+i} = \sum_{j=1}^{n} p_j q_{ij} \quad , \quad i = 1, \ldots, m \quad , \quad \mathcal{C} = \sum_{i=1}^{n+m} p_i c_i \quad , \quad \frac{\partial \mathcal{C}}{\partial q_{uv}} = p_v c_{n+u}$$

† A little more complicated definition (using almost general diagonal matrix **P**) enables the correct description of the dependent variable no matter what linear transformation of the matrix **q** is actually in use. Moreover, dynamic definition of the analytical concentration $\mathcal{C}$ of the "active" component (**not** exclusively reactant or complex) is <u>necessary</u> for the calculation of unknown chemical parameters **c**, **n**, V, **c**', **q** (the value of $\mathcal{C}$ changes).

$$A_{uv} = p_v c_i \left(1 - y\right)/\mathcal{C} \quad , \quad n + u = i \quad , \quad v = 1, \ldots, n$$

$$A_{uv} = -p_v c_{n+u} y/\mathcal{C} \quad , \quad u = 1, \ldots, m \ \wedge \ n + u \neq i \quad , \quad v = 1, \ldots, n$$

Note: Function of type -6 describes the dependence of an intensive quantity y (intensity of an isolated (separated) peak in some "slow exchange" spectrometry, usually NMR spectroscopy, intensity of an isolated peak in capillary zone electrophoresis of an inert system etc., normalized to the sum of intensities of all the peaks embracing all components containing the active component) on the free concentrations $\boldsymbol{c}$ of all components (but mainly on c_i).

1.3 Regression model

Let m is the number of points, i.e. the number of ordered pairs of measured variables on abscissa and ordinate $[x, y]$ ($[x_i, y_i]$, $i = 1, \ldots, m$) and n is the number of unknown parameters. If $m = n$, the values of the unknown parameters are the unique solution of a system of equations. If $m > n$, the system of equations is over-determined. Moreover, if the values of variables x, y and *known* parameters $\mathbf{p} \equiv \boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}$ are accompanied with measuring errors, the solution of the system of equations is **not** unique.

Let the measuring errors $\varepsilon_x, \varepsilon_y, \varepsilon_{p_k}$ of variables x, y, p_k are random, independent, have normal distributions, zero means, variances $\sigma_x^2, \sigma_y^2, \sigma_{p_k}^2$ and act additively. Then the variables x, y, p_k are random, independent, have normal distributions, unknown means and variances $\sigma_x^2, \sigma_y^2, \sigma_{p_k}^2$

$$E(\varepsilon_x) = E(\varepsilon_y) = E(\varepsilon_{p_k}) = 0 \ ,$$
$$E(\varepsilon_x \varepsilon_y) = E(\varepsilon_x) E(\varepsilon_y) = 0 \ ,$$
$$E(\varepsilon_x \varepsilon_{p_k}) = E(\varepsilon_x) E(\varepsilon_{p_k}) = 0 \ ,$$
$$E(\varepsilon_y \varepsilon_{p_k}) = E(\varepsilon_y) E(\varepsilon_{p_k}) = 0 \ ,$$
$$E(\varepsilon_{p_k} \varepsilon_{p_l}) = E(\varepsilon_{p_k}) E(\varepsilon_{p_l}) = 0 \quad , \quad k \neq l \ ,$$
$$E(\varepsilon_x \varepsilon_y) = E(\varepsilon_x \varepsilon_{p_k}) = E(\varepsilon_y \varepsilon_{p_k}) = E(\varepsilon_{p_k} \varepsilon_{p_l}) = 0 \quad , \quad k \neq l \ .$$

However, it is usually assumed that only the values of the variables x and p_k are set (measured) independently in the course of a titration. Since it still holds that $y = f(\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}, x) \equiv f(\mathbf{p}, x)$, an error of the functional relation $f(\mathbf{p}, x)$ should be added to the "independent" measuring error ε_y of the variable y to form the true regression error ε.

From the linear Taylor approximation of the function $f(\mathbf{p}, x)$ in a close neighborhood of the means of the variables x and p_k it follows that

$$\varepsilon \doteq \frac{\partial y}{\partial x} \varepsilon_x + \varepsilon_y + \sum_k \frac{\partial y}{\partial p_k} \varepsilon_{p_k} \P \ ,$$

¶ all the necessary derivatives are given, except for

$$\frac{\partial \mathrm{C}_k}{\partial x} = \frac{(\mathrm{c}'_k - \mathrm{c}_k)\mathrm{V}}{(\mathrm{V} + x)^2} = \frac{\mathrm{c}'_k \mathrm{V} - \mathrm{n}_k}{(\mathrm{V} + x)^2}$$

$$E(\varepsilon) \doteq 0 \quad , \quad D(\boldsymbol{\varepsilon}) \doteq \mathbf{V} \ .$$

The elements of the variance-covariance matrix $\mathbf{V}$ are¶

$$v_{ii} \doteq D(\varepsilon_i) \doteq E(\varepsilon_i^2) \ ,$$

$$v_{ii} = \left(\frac{\partial y}{\partial x}\right)^2_{x=x_i, y=y_i} \sigma^2_{x_i} + \sigma^2_{y_i} + \sum_k \left(\frac{\partial y}{\partial p_k}\right)^2_{y=y_i, p_k=p_{ki}} \sigma^2_{p_{ki}} \ ,$$

$$v_{ij} \doteq cov(\varepsilon_i, \varepsilon_j) \doteq E(\varepsilon_i \varepsilon_j) \ ,$$

$$E(\varepsilon_{x_i}\varepsilon_{x_j}) = E(\varepsilon_{x_i})E(\varepsilon_{x_j}) = 0 \quad , \quad x_i \not\equiv x_j \ ,$$

$$E(\varepsilon_{x_i}\varepsilon_{x_j}) = \sigma^2_{x_i} \quad , \quad x_i \equiv x_j \ ,$$

$$E(\varepsilon_{y_i}\varepsilon_{y_j}) = E(\varepsilon_{y_i})E(\varepsilon_{y_j}) = 0 \quad , \quad i \neq j \ ,$$

$$E(\varepsilon_{p_{ki}}\varepsilon_{p_{kj}}) = E(\varepsilon_{p_{ki}})E(\varepsilon_{p_{kj}}) = 0 \quad , \quad p_{ki} \not\equiv p_{kj} \ ,$$

$$E(\varepsilon_{p_{ki}}\varepsilon_{p_{kj}}) = \sigma^2_{p_{ki}} \quad , \quad p_{ki} \equiv p_{kj} \ ,$$

$\forall i, j, k : i \neq j \ \wedge \ x_i \not\equiv x_j \ \wedge \ p_{ki} \not\equiv p_{kj}$

$$v_{ij} = 0 \ ,$$

$\forall i, j, k : i \neq j \ \wedge \ x_i \equiv x_j \ \wedge \ p_{ki} \equiv p_{kj}$

$$v_{ij} = \left(\frac{\partial y}{\partial x}\right)_{x=x_i, y=y_i} \left(\frac{\partial y}{\partial x}\right)_{x=x_j, y=y_j} \sigma^2_{x_i} + \sum_k \left(\frac{\partial y}{\partial p_k}\right)_{y=y_i, p_k=p_{ki}} \left(\frac{\partial y}{\partial p_k}\right)_{y=y_j, p_k=p_{kj}} \sigma^2_{p_{ki}} \ .$$

If the "independent" measuring error $\boldsymbol{\varepsilon}_y$ is a predominant part of the true regression error $\boldsymbol{\varepsilon}$, the distribution of the random vector $\boldsymbol{\varepsilon}$ is close to normal and the maximum likelihood estimate of the unknown parameters is close to the solution of the over-determined set of m equations (for n unknown variables) minimizing the generalized least squares method criterion§.

All the elements of the vector $\boldsymbol{\vartheta}$ of all the calibration functions implemented so far are "linear", and thus the OPIUM program does not require their initial estimates. When selected elements of the vector $\boldsymbol{\vartheta}$ are the only unknown parameters, then a calculation is made of their unbiased estimates by the iterative

¶ indexing of variables was taken into account, i.e.

$$\varepsilon_i = \left(\frac{\partial y}{\partial x}\right)_{x=x_i, y=y_i} \varepsilon_{x_i} + \varepsilon_{y_i} + \sum_k \left(\frac{\partial y}{\partial p_k}\right)_{y=y_i, p_k=p_{ki}} \varepsilon_{p_{ki}}$$

§ In addition to a point estimate of mean of estimate of the unknown parameter(s) the OPIUM program also offers classical analysis of this estimate in a linearization approximation (variance, t- and F-interval estimate of mean, significance test, correlation matrix, partial and multiple correlation coefficients etc.) as well as jack-knife and likelihood ratio analysis (interval estimate of mean, significance test). At each point, many types of residuals are calculated along with a number of scalar and vector quantities for indication of golden points, outliers and high leverage points. The OPIUM program also tests several hypotheses on homogeneity, independence and normality of the standardized residuals.

linear least squares method (the variance-covariance matrix $\mathbf{V}$ is in general a function of the unknown parameters), where biased estimates are also available, obtained by the ridge regression method.

1.4 Principal component analysis

The dependence of the measured variable y on the free concentrations $\boldsymbol{c}$ of components in a common spectrophotometric or "fast exchange" NMR experiment¶ can be expressed in a matrix form

$$\mathbf{A} = \boldsymbol{\varepsilon}\,\boldsymbol{C} \quad , \quad \mathbf{A} = \boldsymbol{\delta}\boldsymbol{\Pi}\boldsymbol{C}/\mathcal{C}\ \dagger,$$

$$\mathbf{A} \sim (m,n) \quad , \quad \boldsymbol{\varepsilon} \sim (m,k) \quad , \quad \boldsymbol{\delta} \sim (m,k) \quad , \quad \boldsymbol{C} \sim (k,n) \ ,$$

m is the number of wavelengths (NMR active nuclei), n is the number of solutions, k is the number of independent ("active") components, $\mathbf{A}$ is the absorbance matrix (the matrix of chemical shifts) ($a_{ij} \equiv y_{(i-1)\times n+j}$), $\boldsymbol{\varepsilon}$ is the matrix of molar absorption coefficients, $\boldsymbol{\delta}$ is the matrix of the chemical shifts of pure components and $\boldsymbol{C}$ is the matrix of the free concentrations of components (a column per solution).

Obviously $r(\mathbf{A}) \leq \min(m,n,k)$. Thus, if both $m > k$ and $n > k$, the true rank of the absorbance matrix (the matrix of chemical shifts) gives us a reasonable estimate of the number of light absorbing (δ influencing) components in a solution $k \geq r(\mathbf{A})$. Moreover, the number of observations can dramatically be reduced from $m \times n$ to $k \times n$ by means of the singular value decomposition of the matrix $\mathbf{A}$ (no more than k suitable linear combinations of the m original wavelengths are needed).

For $m \geq n$

$$\mathbf{A} = \mathbf{U}_n^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{V} \ ,$$

$$\mathbf{U}_n \sim (n,m) \quad , \quad \boldsymbol{\Sigma} \sim (n,n) \quad , \quad \mathbf{V} \sim (n,n) \ ,$$

$$\mathbf{U}_n\mathbf{U}_n^{\mathrm{T}} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{1}_n \ ,$$

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1,\ldots,\sigma_n) \quad , \quad \sigma_1 \geq \ldots \geq \sigma_n \geq 0 \ ,$$

$$\boldsymbol{\Sigma}\mathbf{V} = \mathbf{U}_n\boldsymbol{\varepsilon}\boldsymbol{C} \quad , \quad \boldsymbol{\Sigma}\mathbf{V} = \mathbf{U}_n\boldsymbol{\delta}\boldsymbol{\Pi}\boldsymbol{C}/\mathcal{C} \ .$$

Since $r(\mathbf{A}) \doteq k$, $\boldsymbol{\Sigma} \doteq \mathrm{diag}(\sigma_1,\ldots,\sigma_k,0_{k+1},\ldots,0_n)$§. Then

$$\boldsymbol{\Sigma}_k\mathbf{V}_k \doteq \mathbf{U}_k\boldsymbol{\varepsilon}\boldsymbol{C} \quad , \quad \boldsymbol{\Sigma}_k\mathbf{V}_k \doteq \mathbf{U}_k\boldsymbol{\delta}\boldsymbol{\Pi}\boldsymbol{C}/\mathcal{C} \ ,$$

¶ see the definitions of calibration functions of types 7 and -5

† Since only the "active" components are taken into account, matrix $\boldsymbol{\Pi}$ is diagonal containig only the non-zero diagonal elements of the matrix $\mathbf{P}$ (see the definition of a calibration function of type -5), i.e. $\boldsymbol{\Pi} \sim (k,k)$, $r(\boldsymbol{\Pi}) = r(\mathbf{P}) = k$.

§ To estimate the "effective" rank k of the matrix $\mathbf{A}$ (both the equations would hold exactly only if where the matrices $\mathbf{\Sigma}_k \sim (k,k)$, $\mathbf{V}_k \sim (k,n)$ and $\mathbf{U}_k \sim (k,m)$ are appropriate submatrices of the matrices $\mathbf{\Sigma}$, $\mathbf{V}$ and $\mathbf{U}_n$¶.

For $m < n$

$$\mathbf{A} = \mathbf{U}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{V}_m \ ,$$

$$\mathbf{U} \sim (m,m) \quad , \quad \mathbf{\Sigma} \sim (m,m) \quad , \quad \mathbf{V}_m \sim (m,n) \ ,$$

$$\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{V}_m\mathbf{V}_m^{\mathrm{T}} = \mathbf{1}_m \ ,$$

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_m) \quad , \quad \sigma_1 \geq \ldots \geq \sigma_m \geq 0 \ ,$$

$$\mathbf{\Sigma}\mathbf{V}_m = \mathbf{U}\boldsymbol{\varepsilon}\boldsymbol{C} \quad , \quad \mathbf{\Sigma}\mathbf{V}_m = \mathbf{U}\boldsymbol{\delta}\mathbf{\Pi}\boldsymbol{C}/\mathcal{C} \ .$$

Since $r(\mathbf{A}) \doteq k$, $\mathbf{\Sigma} \doteq \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0_{k+1}, \ldots, 0_m)$. Then

$$\mathbf{\Sigma}_k\mathbf{V}_k \doteq \mathbf{U}_k\boldsymbol{\varepsilon}\boldsymbol{C} \quad , \quad \mathbf{\Sigma}_k\mathbf{V}_k \doteq \mathbf{U}_k\boldsymbol{\delta}\mathbf{\Pi}\boldsymbol{C}/\mathcal{C} \ ,$$

where the matrices $\mathbf{\Sigma}_k \sim (k,k)$, $\mathbf{V}_k \sim (k,n)$ and $\mathbf{U}_k \sim (k,m)$ are appropriate submatrices of the matrices $\mathbf{\Sigma}$, $\mathbf{V}_m$ and $\mathbf{U}$.

1.5 "Trial and error" method

Let p, k, l are natural numbers and m, n are non-negative integers ($m+n < p \ \wedge \ k,l > m$) with the following meaning:

p	...	number of "unknown" complexes,
m	...	number of "unknown" complexes with *higher* priority (all of them **must** definitely be in every generated combination),
n	...	number of "unknown" complexes with *lower* priority,
$p-m-n$	...	number of "unknown" complexes with *common* priority (at least one of them **must** be in each generated combination),
k, l	...	lower (upper) limit of order of combination.

Then the number of combinations (chemical model hypotheses) is

$$\begin{array}{rc} n < k-m : & \sum_{i=k-m}^{l-m} \binom{p-m}{i} \\ k-m \leq n \leq l-m : & \sum_{i=k-m}^{l-m} \binom{p-m}{i} - \sum_{i=k-m}^{n} \binom{n}{i} \\ n > l-m : & \sum_{i=k-m}^{l-m} \binom{p-m}{i} - \sum_{i=k-m}^{l-m} \binom{n}{i} \end{array} \ .$$

---

there were no measuring errors), sequence of residual standard deviations is calculated[3]

$$s_i = \sqrt{\frac{\mathrm{tr}\,(\mathbf{\Sigma}^2) - \sum_{j=1}^{i}\sigma_j^2}{p-i}} \quad , \quad i = 0, \ldots, p-1 \quad , \quad p = \min(m,n) \ .$$

Then inequalities $s_0 > \ldots > s_{k-1} > \sigma_y \geq s_k \geq \ldots \geq s_p \geq 0$ should be valid, where σ_y is the standard deviation of the random variable y (absorbance, chemical shift).

¶ i.e. $\mathbf{A}' \doteq \boldsymbol{\varepsilon}'\boldsymbol{C}$, $\mathbf{A}' \doteq \boldsymbol{\delta}'\mathbf{\Pi}\boldsymbol{C}/\mathcal{C}$, where the matrices $\mathbf{A}' = \mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}} \sim (k,n)$, $\boldsymbol{\varepsilon}' = \mathbf{U}_k^{\mathrm{T}}\boldsymbol{\varepsilon} \sim (k,k)$ and $\boldsymbol{\delta}' = \mathbf{U}_k^{\mathrm{T}}\boldsymbol{\delta} \sim (k,k)$ are "reduced" equivalents of the matrices $\mathbf{A}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$

For each combination, the OPIUM program creates a shortened information about the goodness of fit (containing quantities such as sum of squares of residuals (U), Akaike information criterion (AIC), Schwarz[4]-Rissanen[5], Hannan--Quinn[6] and Anděl-Perez-Negrao[7] statistic) and stores it into a file¶.

2 Data Analysis †§

2.1 Potentiometry

$\boxed{\text{2-methylaminomethylpyrrole-4-sulphonic acid } (H_2\mathit{psma}),\ Ag^+}$

The 1st example[8] is the potentiometric titration of 2-methylaminomethylpyrrole-4-sulphonic acid and Ag^+ (of known analytical concentration c_1 of H^+, c_2 of psma^{2-} and c_3 of Ag^+, the last two at the variable molar ratio $1:1$ up to $2:1$) with potassium hydroxide (of known analytical concentration $c_1' < 0$ of H^+). The glass electrode calibrated in concentration scale ($\mathrm{pH} = -\log c_1$) was used and the dependence of $-\log c_1$ on the added volume of the titrant was measured‡. The origin of individual parameters is displayed as an "extension" of the matrix $\mathbf{q}$ (the chemical model is taken from Stünzi H. et al[8]).

H^+	psma^{2-}	Ag^+		
1	0	0	H^+	c_1, c_1'
0	1	0	psma^{2-}	c_2
0	0	1	Ag^+	c_3
−1	0	0	OH^-	β_{-100}
−1	0	1	AgOH	β_{-101}
1	1	0	$H\mathit{psma}^-$	β_{110}
2	1	0	$H_2\mathit{psma}$	β_{210}
1	1	1	$AgH\mathit{psma}$	β_{111}
2	2	1	$AgH_2\mathit{psma}_2^-$	β_{221}
0	2	2	$Ag_2\mathit{psma}_2^{2-}$	β_{022}
1	3	3	$Ag_3H\mathit{psma}_3^{2-}$	β_{133}
0	3	3	$Ag_3\mathit{psma}_3^{3-}$	β_{033}

¶ quite suitable for processing with standard Unix tools (`grep`, `sort`, `awk` etc.)

† common units are used in the rest of the text (and will no more be mentioned)

quantity	*unit*
mass amount	mmol
volume	ml
concentration	$\mathrm{mol \cdot l^{-1}}$

§ Unit variance-covariance matrix $\mathbf{V} = \mathbf{1}$ is used in all the calculations to enable the comparison of the estimates found by different programs. On this assumption, all the programs using the same least squares criterion should give absolutely the same results, i.e. point estimate(s) of mean(s) of the estimate(s) of the unknown parameter(s).

‡ 1 calibration function of type −1 is used

The results are presented in *Table 1.*

	HYPERQUAD[9]	OPIUM
$\log\beta_{111}$	18.00(1)	17.98(1)
$\log\beta_{221}$	36.45(1)	36.43(1)
$\log\beta_{022}$	20.17(1)	20.12(2)
$\log\beta_{133}$	41.01(2)	41.00(2)
$\log\beta_{033}$	31.39(3)	31.40(2)
$10^4 \cdot U$	115.4	115.4

Table 1.

When applying "trial and error" method, the ability of the program to get the estimates of the unknown parameters even from their bad initial estimates can save much time¶. That is why the robustness of both the programs used is tested before the best model is found. The results are presented in *Table 2.*

initial estimates	HYPERQUAD	OPIUM
20, 35, 20, 40, 30	+	+
10, 30, 10, 30, 20	no convergence [a]	+
20, 35, 20, 0, 0	fails in 23rd iteration [b]	+
0, 0, 0, 40, 30	+	+
0, 0, 0, 0, 0	no convergence [a]	+
20, 35, 20, 50, 40	+	+
30, 30, 30, 40, 30	fails in 174th iteration [b]	+
25, 30, 25, 35, 35	fails in 10th iteration [b]	+
15, 40, 15, 45, 25	+	+
10, 45, 10, 50, 20	+	+
20, 20, 20, 20, 20	fails in 24th iteration [b]	+
50, 50, 50, 50, 50	fails in 2nd iteration [b]	+ [d]
0, 50, 0, 50, 0	fails in 6th iteration [b]	+
50, 0, 50, 0, 50	fails in 1st iteration [c]	+ [d]
50, 50, 50, 0, 0	fails in 1st iteration [c]	+ [d]
0, 0, 0, 50, 50	+	+

[a] possibly a very very slow convergence
[b] MS DOS© collapses, rebooting is necessary
[c] floating point error, program exits normally
[d] takes approximately 5 minutes on i486DX 33MHz

Table 2.

¶ not to mention the possibility of passing over an existing solution

To determine the best model, we chose 23 "possible" complexes (corresponding rows of the matrix of stoichiometric coefficients $\mathbf{q}$ (H^+, $psma^{2-}$, Ag^+) are

	3. 2 2 1	6. 2 2 2	9. 2 3 2	12. 2 3 3	15. 2 4 2	18. 2 4 3	21. 2 4 4
1. 1 1 1	4. 1 2 1	7. 1 2 2	10. 1 3 2	13. 1 3 3	16. 1 4 2	19. 1 4 3	22. 1 4 4
2. 0 1 1	5. 0 2 1	8. 0 2 2	11. 0 3 2	14. 0 3 3	17. 0 4 2	20. 0 4 3	23. 0 4 4

and obtained $\sum_{i=1}^{5} \binom{23}{i} = 44551$ combinations of order 1 up to 5¶.

To demonstrate the power of the OPIUM program, a "brute force" approach was used and <u>all the combinations</u> were tested†. 20818 solutions were found, but only 4035 of them are "positive" (i.e. point estimates of all the unknown stability constants are positive)§.

```
1. {5} o.k. 1.1539e-2 -1568.7135 -9.3825 -9.4384 1.7273e-4 "agpsma +[1 3 8 13 14]"
2. {5} o.k. 1.6055e-2 -1514.2158 -9.0522 -9.1081 2.4033e-4 "agpsma +[1 3 8 14 22]"
3. {5} o.k. 1.6485e-2 -1509.8532 -9.0257 -9.0816 2.4677e-4 "agpsma +[1 3 8 13 23]"
4. {5} o.k. 1.8080e-2 -1494.6209 -8.9334 -8.9893 2.7064e-4 "agpsma +[1 3 8 22 23]"
5. {5} o.k. 2.3136e-2 -1453.9293 -8.6868 -8.7427 3.4633e-4 "agpsma +[1 3 8 14 18]"
.
.
13. {4} o.k. 3.7961e-2 -1374.2290 -8.2288 -8.2736 4.9893e-4 "agpsma +[1 3 8 22]"
.
.
19. {4} o.k. 4.4369e-2 -1348.4928 -8.0728 -8.1176 5.8315e-4 "agpsma +[1 3 8 13]"
.
.
4035. {1} o.k. 150.8192 -12.8275 -5.2839e-2 -6.4022e-2 1.1762 "agpsma +[6]"
```

As you can see, the best model is really the one analyzed above. For comparison, several "worse" alternatives are also presented in *Table 3*.

Some attempts to improve the goodness of fit and to confirm the determined model ($\mathbf{q}$, $\boldsymbol{\beta}$) by the ESI approach[2] are presented in *Table 4*. The linear parameter E_0‡ represents the systematic error of the pH measurements here and should be 0, the original values of c_2 (the analytical concentrations of $psma^{2-}$) are 0.01001, 0.0203, 0.005, 0.00202, 0.01005, 0.00303, 0.00404¶, the correct values of q_{53}, q_{63}, q_{73} (the stoichiometric coefficients of Ag^+ in the first three "unknown" complexes, i.e. 0 1 1, 2 2 1, 0 2 2) are 1, 1, 2.

¶ if some preliminary analysis of the data had been done, the number of combinations could have been reduced substantially[8]

† the task was run overnight on Alpha 21164A 300MHz, which is about 20 × faster than i486DX 33MHz

§ each line in the table gives a shortened information about respective model and goodness of fit: `5` is the order of the combination (i.e. the number of unknown complexes), `o.k.` means a positive solution (`'N'` would mean a negative solution), sum of squares of residuals (U), Akaike information criterion (AIC), Schwarz-Rissanen, Hannan-Quinn (1), Anděl-Perez-Negrao (1, $\frac{1}{4}$) statistic, `agpsma` is the name of the input file being processed and `+[1 3 8 13 14]` is the set of complexes forming the model (the quoted information is in the format required by the OPIUM program as the command line parameter to "rerun" the model and produce the complete output file)

‡ 7 calibration functions of type 2 are used, $S = -1$

¶ the E_0 and c_2 parameters are supposed to be individual properties of each of the 7 titration curves

	1.	2.	3.	4.	5.	13.	19.
$\log\beta_{111}$	17.98(1)	17.93(1)	18.00(1)	17.95(1)	18.00(1)	17.96(2)	18.00(2)
$\log\beta_{221}$	36.43(1)	36.50(1)	36.40(1)	36.48(1)	36.43(1)	36.46(1)	36.39(2)
$\log\beta_{022}$	20.12(2)	20.17(2)	20.23(1)	20.246(9)	20.13(2)	20.339(6)	20.350(7)
$\log\beta_{133}$	41.00(2)	-	41.02(2)	-	-	-	40.78(3)
$\log\beta_{033}$	31.40(2)	31.27(3)	-	-	31.39(4)	-	-
$\log\beta_{243}$	-	-	-	-	59.37(3)	-	-
$\log\beta_{144}$	-	52.39(2)	-	52.42(2)	-	52.21(3)	-
$\log\beta_{044}$	-	-	42.55(3)	42.41(4)	-	-	-
$10^4 \cdot U$	115.4	160.6	164.9	180.8	231.4	379.6	443.7
AIC	-1569	-1514	-1510	-1495	-1454	-1374	-1348

Table 3.

	E_0, $\boldsymbol{\beta}$	$\mathbf{c}$, $\boldsymbol{\beta}$	$\mathbf{q}$, $\boldsymbol{\beta}$	E_0, $\mathbf{c}$, $\boldsymbol{\beta}$	$\mathbf{c}$, $\mathbf{q}$, $\boldsymbol{\beta}$	boundary condition
E_0	$-0.006(6)$	·	·	0.005(5)	·	-
E_0	0.008(7)	·	·	$-0.005(5)$	·	-
E_0	0.007(6)	·	·	0.002(5)	·	-
E_0	0.007(5)	·	·	$-0.007(4)$	·	-
E_0	0.007(4)	·	·	0.012(5)	·	-
E_0	0.015(5)	·	·	0.010(4)	·	-
E_0	0.010(4)	·	·	0.019(4)	·	-
$10^2 \cdot c_2$	·	1.006(1)	·	1.005(1)	1.003(2)	$c_1 = 0.01967 + 2c_2$
$10^2 \cdot c_2$	·	2.003(3)	·	1.991(2)	1.998(4)	$c_1 = 0.01987 + 2c_2$
$10^3 \cdot c_2$	·	4.974(7)	·	4.963(7)	4.96(1)	$c_1 = 0.01975 + 2c_2$
$10^3 \cdot c_2$	·	2.018(4)	·	1.991(5)	2.014(7)	$c_1 = 0.00398 + 2c_2$
$10^2 \cdot c_2$	·	0.999(1)	·	1.007(2)	0.998(3)	$c_1 = 0.01970 + 2c_2$
$10^3 \cdot c_2$	·	2.996(5)	·	3.015(7)	2.99(1)	$c_1 = 0.00398 + 2c_2$
$10^3 \cdot c_2$	·	4.015(6)	·	4.09(1)	4.01(2)	$c_1 = 0.00398 + 2c_2$
q_{53}	·	·	0.99(2)	·	0.99(1)	-
q_{63}	·	·	1.06(2)	·	1.039(6)	-
q_{73}	·	·	2.048(9)	·	2.01(2)	-
$\log\beta_{111}$	17.99(1)	17.984(4)	17.94(4)	17.963(8)	17.95(2)	-
$\log\beta_{221}$	36.44(1)	36.442(5)	36.58(4)	36.48(1)	36.55(2)	-
$\log\beta_{022}$	20.21(3)	20.11(2)	20.31(4)	20.13(1)	20.13(5)	-
$\log\beta_{133}$	40.97(4)	40.996(9)	40.95(2)	40.97(3)	41.00(1)	-
$\log\beta_{033}$	31.30(7)	31.49(3)	31.36(3)	31.51(5)	31.53(2)	-
$10^4 \cdot U$	50.80	17.84	94.76	6.018	12.29	
AIC	-1690	-1863	-1595	-2028	-1918	

· means the original (correct) value

Table 4.

2.2 Spectrophotometry

methionine (H*met*), SCN^-, Pd^{2+}

The 2nd example[10,11] is the spectrophotometric "titration" of methionine¶, SCN^- and Pd^{2+} at the fixed molar ratio 2 : 1000 : 1 and pH 0 up to 11. The free concentration c_1 of H^+ (i.e. pH) and the analytical concentration C_2 of met^-, C_3 of SCN^- and C_4 of Pd^{2+} are known in each of the 30 solutions. The dependence of the absorbance at 25 wavelengths (260 up to 380 nm) on the free concentration c_1 (for fixed analytical concentrations C_2, C_3 and C_4) was measured†. The origin of individual parameters is displayed as an "extension" of the matrix $\mathbf{q}$ (the chemical model is taken from Stypinski B.[10]).

H^+	met^-	SCN^-	Pd^{2+}		
1	0	0	0	H^+	$30 \times c_1$
0	1	0	0	met^-	$30 \times C_2$
0	0	1	0	SCN^-	$30 \times C_3$
0	0	0	1	Pd^{2+}	$30 \times C_4$
1	1	0	0	$Hmet$	β_{1100}
2	1	0	0	H_2met^+	β_{2100}
0	0	4	1	$Pd(SCN)_4^{2-}$	$25 \times a_3$, β_{0041}
0	1	2	1	$Pdmet(SCN)_2^-$	$25 \times a_4$, β_{0121}
0	2	0	1	$Pdmet_2$	$25 \times a_5$, β_{0201}

Principal component analysis gives the set of residual standard deviations of absorbance 0.2581, 0.0466, 0.035, 0.0009, i.e. $k \approx 3$ (there are at least 3 light absorbing species). The results are presented in *Table 5.*

	SPECFIT[12]	OPIUM
$\log \beta_{0121}$	31.561(7)	31.561(8)
$\log \beta_{0201}$	34.33(2)	34.33(3)
$10^4 \cdot U$	98.96	104.9

Table 5.

When only 3 principal components of the absorbance matrix are selected, the OPIUM program gives the sum of squares of residuals $10^4 \cdot U = 98.96$ (which results in smaller values of the estimated standard deviations).

2-methyl-3-hydroxy-4H-pyran-4-one (maltol, *maltol*), UO_2^{2+}

The 3rd example[13] is the spectrophotometric "titration" of maltol and UO_2^{2+}

¶ 2-amino-4-(methylthio)butanoic acid

† 25 calibration functions of type 7 are used

at the variable molar ratio 10 : 1 up to 300 : 1 and pH 1 up to 9. The free concentration c_1 of H^+ (i.e. pH) and the analytical concentration C_2 of *maltol* and C_3 of UO_2^{2+} are known in each of the 148 solutions. The dependence of the absorbance at 3 wavelengths (390, 400 and 410 nm using 2 different optical lengths) on the free concentration c_1 and the analytical concentrations C_2 and C_3 was measured¶. The origin of individual parameters is displayed as an "extension" of the matrix $\mathbf{q}$ (the chemical model is taken from Havel J.[13]).

H^+	*maltol*	UO_2^{2+}		
1	0	0	H^+	$148 \times c_1$
0	1	0	*maltol*	$148 \times C_2$
0	0	1	UO_2^{2+}	$148 \times C_3$
−1	1	0	$H_{-1}maltol^-$	β_{-110}
0	1	1	$UO_2maltol^{2+}$	$3 \times a_2,\ \beta_{011}$
−1	1	1	$UO_2H_{-1}maltol^+$	$3 \times a_3,\ \beta_{-111}$
−1	2	1	$UO_2H_{-1}maltol_2^+$	$3 \times a_4,\ \beta_{-111}$
−2	2	1	$UO_2H_{-2}maltol_2$	$3 \times a_5,\ \beta_{-221}$
−2	3	1	$UO_2H_{-2}maltol_3$	$3 \times a_6,\ \beta_{-231}$
−3	3	1	$UO_2H_{-3}maltol_3^-$	$3 \times a_7,\ \beta_{-331}$
−3	1	1	$UO_2H_{-3}maltol^-$	$3 \times a_8,\ \beta_{-311}$
−6	4	2	$(UO_2)_2H_{-6}maltol_4^{2-}$	$3 \times a_9,\ \beta_{-642}$
−6	3	2	$(UO_2)_2H_{-6}maltol_3^{2-}$	$3 \times a_{10},\ \beta_{-632}$
−6	1	2	$(UO_2)_2H_{-6}maltol^{2-}$	$3 \times a_{11},\ \beta_{-612}$

In spite of the enormous number of solutions, the absorbance data do not contain enough information for computation of estimates of all the unknown parameters ($\mathbf{a}$, $\boldsymbol{\beta}$) together (and no "overall" calculation has ever been performed). The 10 unknown complexes were determined by dividing the studied pH range into several parts and treating them separately.

We did not follow this procedure, but searched for the best "overall" model. For the sake of simplicity, we chose the 10 already "determined" complexes and obtained $\sum_{i=1}^{10} \binom{10}{i} = 1023$ combinations of order 1 up to 10.

There was no problem to test all the combinations†. 740 solutions were found, 535 of them are "positive" (i.e. point estimates of all the unknown stability constants are positive), but most of them are highly correlated. That is why only positive solutions of orders 1 up to 5 are listed (421 solutions).

```
1. {5} o.k. 1.5468e-2 -4517.5667 -9.9441 -10.0559 1.9543e-4 "uo2malt +10 9 5 3 1"
2. {5} o.k. 1.6792e-2 -4481.1064 -9.8620 -9.9737 2.1216e-4 "uo2malt +10 7 5 3 1"
3. {5} o.k. 1.7515e-2 -4462.3811 -9.8198 -9.9316 2.2130e-4 "uo2malt +8 7 5 3 1"
```

¶ 3 calibration functions of type 8 are used

† the task took several minutes on R10000 175MHz, which is about 20 × faster than i486DX 33MHz

```
4. {5} o.k. 1.7573e-2 -4460.9123 -9.8165 -9.9283 2.2203e-4 "uo2malt +10 8 5 3 1"
5. {5} o.k. 2.0239e-2 -4398.2185 -9.6753 -9.7870 2.5570e-4 "uo2malt +9 7 5 3 1"
6. {5} o.k. 2.1360e-2 -4374.2718 -9.6214 -9.7331 2.6987e-4 "uo2malt +9 8 5 3 1"
7. {5} o.k. 2.6417e-2 -4279.9353 -9.4089 -9.5206 3.3376e-4 "uo2malt +10 9 7 5 3"
8. {4} o.k. 2.7340e-2 -4272.6845 -9.4389 -9.5283 2.8653e-4 "uo2malt +[2 4 6 8]"
9. {5} o.k. 3.3617e-2 -4172.9169 -9.1679 -9.2796 4.2473e-4 "uo2malt +10 9 8 3 1"
10. {5} o.k. 3.4558e-2 -4160.6572 -9.1403 -9.2520 4.3662e-4 "uo2malt +10 9 8 5 3"
11. {4} o.k. 3.5413e-2 -4157.8077 -9.1801 -9.2695 3.7114e-4 "uo2malt +[2 4 6 7]"
.
.
421. {1} o.k. 12.6952 -1570.2417 -3.4906 -3.5130 5.3995e-2 "uo2malt +[1]"
```

The best models (of orders 5 and 4) are presented in *Table 6.*

	1.	2.	3.	5.	8.	11.
$\log\beta_{-111}$	$-0.40(2)$	$-0.40(2)$	$-0.39(2)$	$-0.39(2)$	$-0.38(2)$	$-0.40(3)$
$\log\beta_{-221}$	$-2.33(3)$	$-2.31(4)$	$-2.30(4)$	$-2.28(4)$	$-2.25(5)$	$-2.33(5)$
$\log\beta_{-331}$	$-7.57(4)$	$-7.59(4)$	$-7.49(4)$	$-7.50(5)$	$-7.42(5)$	$-7.48(6)$
$\log\beta_{-311}$	$-12.20(6)$	-	-	-	-	$-11.92(6)$
$\log\beta_{-642}$	$-15.49(8)$	$-15.1(1)$	-	$-15.45(9)$	$-15.6(1)$	-
$\log\beta_{-632}$	-	$-17.4(1)$	$-17.17(9)$	-	-	-
$\log\beta_{-612}$	-	-	$-22.7(1)$	$-23.8(1)$	-	-
$10^4 \cdot U$	154.7	167.9	175.2	202.4	273.4	354.1
AIC	-4518	-4481	-4462	-4398	-4273	-4158

Table 6.

2.3 "Fast exchange" NMR

4,9-diaza-1,12-dodecanediamine (spermine, *sprm*)

The 4th example[14,15] is the ^{1}H+^{13}C NMR "titration" of spermine. The free concentration c_1 of H^+ (i.e. pH) and the analytical concentration C_2 of *sprm* are known in each of the 20 solutions. There are 10 different types of nuclei (5×H, 5×C) in the molecule and thus the dependence of 10 chemical shifts on the free concentration c_1 was measured¶. The origin of individual parameters is displayed as an "extension" of the matrix $\mathbf{q}$.

H^+	*sprm*		
1	0	H^+	$20 \times c_1$
0	1	*sprm*	$10 \times a_{-2}$, $20 \times C_2$
1	1	H*sprm*$^+$	$10 \times a_1$, β_{11}
2	1	H_2*sprm*$^{2+}$	$10 \times a_2$, β_{21}
3	1	H_3*sprm*$^{3+}$	$10 \times a_3$, β_{31}
4	1	H_4*sprm*$^{4+}$	$10 \times a_4$, β_{41}

Principal component analysis gives the set of residual standard deviations

¶ 10 calibration functions of type -5 are used of chemical shift 28.51, 0.8013, 0.0522, 0.0252, 0.0017, 0.0009, i.e. $k \approx 5$ (there are at least 5 δ influencing species, but even for $k = 1$ reliable estimates can be obtained). The results are presented in *Table 8.*

	HYPNMR[14,15]	OPIUM
$\log \beta_{11}$	11.20(2)	11.20(2)
$\log \beta_{21}$	21.49(6)	21.49(6)
$\log \beta_{31}$	30.53(8)	30.53(8)
$\log \beta_{41}$	38.7(1)	38.7(1)
$10^4 \cdot U$	37.09	37.09

Table 8.

The NMR chemical shift(s) of a pure weak acid in a buffered solution is independent on the concentration of the acid¶. That is why no of the 20 analytical concentrations C_2 can be "marked" *unknown.* Formally, it is possible to calculate any of the 20 free concentrations c_1 of H^+, but it can hardly be of any practical importance. Moreover, only a few of the free concentrations can be calculated because of the strong correlation of all the estimates†.

$\boxed{\text{2-aminoethylphosphonic acid } (aep^{2-})}$

The 5th example[16,17] is the ^{31}P NMR titration of 2-aminoethylphosphonic acid (of known mass amount n_1 of H^+ and n_2 of aep^{2-}) with tetramethylammonium hydroxide (of known analytical concentration $c_1' < 0$ of H^+). There is one phosphorus nucleus in the molecule and thus the dependence of one chemical shift on the added volume of the titrant was measured§. The origin of individual parameters is displayed as an "extension" of the matrix **q**.

H^+	aep^{2-}		
1	0	H^+	n_1, c_1'
0	1	aep^{2-}	a_{-2}, n_2
−1	0	OH^-	β_{-10}
1	1	$Haep^-$	a_2, β_{11}
2	1	H_2aep	a_3, β_{21}
3	1	H_3aep^+	a_4, β_{31}

The hypersurface of the sum of squares of residuals with respect to the parameters **a**, $\boldsymbol{\beta}$ is "flat" (in the 8-dimensional space) because of the strong

¶ all the analytical concentrations $C_2 \equiv \mathcal{C}$ in the input file could be replaced by random numbers and the results, i.e. estimates of the unknown parameters, would be the same

† the number of solutions must be substantially greater than the number of unknown parameters

§ 1 calibration function of type −5 is used correlation of the estimates of parameters a_4, β_{21} and β_{31}. At least one of the parameters **n**, V, **c**' has to be "marked" *unknown* to make the problem solvable. Several choices¶ are presented in *Table 9.* The linear parameters a_{-2}, a_2, a_3 and a_4 are the ^{31}P NMR chemical shifts of pure components aep^{2-}, $Haep^{-}$, H_2aep and H_3aep^{+}, the original values of n_1 and n_2 (the mass amounts of H^+ and aep^{2-} in the titrand) are 2.6371 and 0.8760, the original value of c'_1 (the analytical concentration of H^+ in the titrant) is -0.09912.

	a, n_1, $\boldsymbol{\beta}$	**a**, n_2, $\boldsymbol{\beta}$	**a**, c'_1, $\boldsymbol{\beta}$	**a**, **n**, $\boldsymbol{\beta}$	**a**, n_1, c'_1, $\boldsymbol{\beta}$	**a**, n_2, c'_1, $\boldsymbol{\beta}$
a_{-2}	20.16(1)	20.18(1)	20.15(1)	20.17(1)	20.17(1)	20.17(1)
a_2	16.813(1)	16.812(1)	16.815(1)	16.812(1)	16.812(1)	16.812(1)
a_3	19.272(1)	19.263(2)	19.278(1)	19.265(2)	19.265(2)	19.265(2)
a_4	25(1)	23.4(5)	28(2)	23.7(6)	23.7(6)	23.7(6)
n_1	2.6430(3)	·	·	2.639(1)	2.651(2)	·
n_2	·	0.8708(3)	·	0.872(1)	·	0.8715(7)
c'_1	·	·	−0.09879(2)	·	−0.0996(1)	−0.09906(5)
$\log\beta_{11}$	11.361(6)	11.370(5)	11.356(6)	11.367(6)	11.369(6)	11.367(6)
$\log\beta_{21}$	18.71(5)	18.77(5)	18.66(6)	18.76(5)	18.76(5)	18.75(5)
$\log\beta_{31}$	19.7(1)	19.95(8)	19.4(1)	19.9(1)	19.9(1)	19.9(1)
$10^4 \cdot U$	16.31	14.92	19.68	14.76	14.77	14.76
AIC	−1414	−1426	−1391	−1425	−1425	−1425

· means the original value

Table 9.

2.4 "Slow exchange" NMR

1-(2-aminoethylamino)-1-deoxy-D-galacticol (*galac*), Cd^{2+}

The 6th example[18] is the ^{113}Cd NMR "titration" of 1-(2-aminoethylamino)-1-deoxy-D-galacticol and Cd^{2+} at the variable molar ratio 1 : 5 up to 2 : 1 and approximately constant pH ≈ 6.5. The free concentration c_1 of H^+ (i.e. pH) and the analytical concentration C_2 of *galac* and C_3 of Cd^{2+} are known in each of the 9 solutions. There are 3 peaks in the NMR spectrum (assigned to Cd^{2+}, $Cd\,galac^{2+}$ and $Cd\,galac_2^{2+}$) and thus the dependence of three normalized integral intensities on the free concentration c_1 and the analytical concentrations C_2 and C_3 was measured†. The origin of individual parameters is displayed as an "extension" of the matrix **q**.

H^+	*galac*	Cd^{2+}		
1	0	0	H^+	$9 \times c_1$

¶ some of them are a bit obscure, probably the best approach should be the calculation of n_2 using the "natural" boundary condition for n_1, but the information was not available

† 1 calibration function of type −6 is used (with 3 different specifiers i, see the definition of the function)

H^+	*galac*	Cd^{2+}		
0	1	0	*galac*	$9 \times C_2$
0	0	1	Cd^{2+}	$9 \times C_2$
1	1	0	$H\,galac^+$	β_{110}
2	1	0	H_2galac^{2+}	β_{210}
0	1	1	$Cd\,galac^{2+}$	β_{011}
0	2	1	$Cd\,galac_2^{2+}$	β_{021}

The results are presented in *Table 10.*

	Huskens J. et al[18]	OPIUM
$\log \beta_{011}$	*4.6(3)*	4.63(6)
$\log \beta_{021}$	*8.7(4)*	8.6(1)
$10^4 \cdot U$		658.4

Table 10.

^{113}Cd NMR integral intensity data are usually of a poor quality. On the other hand, the chemical shifts of individual peaks give us reliable information about both the number of "active" components and their structure. No matter how imprecise the integral intensities are, they should agree with the proposed chemical model. Such an agreement is tested by the ESI approach and the results are presented in *Table 11.*¶† The correct values of q_{32}, q_{33}, q_{42}, q_{43} (the stoichiometric coefficients of *galac* and Cd^{2+} in both the unknown complexes, i.e. 0 1 1, 0 2 1) are 1, 1, 2, 1.

	q_{32}, $\boldsymbol{\beta}$	q_{33}, $\boldsymbol{\beta}$	q_{42}, $\boldsymbol{\beta}$	q_{43}, $\boldsymbol{\beta}$	q_{32}, q_{42}, $\boldsymbol{\beta}$	q_{33}, q_{43}, $\boldsymbol{\beta}$
q_{32}	0.83(7)	·	·	·	0.88(8)	·
q_{33}	·	1.3(1)	·	·	·	1.3(1)
q_{42}	·	·	3.8(6)	·	2.8(5)	·
q_{43}	·	·	·	0.7(2)	·	0.7(2)
$\log \beta_{011}$	3.8(4)	4.8(1)	5.0(2)	4.8(2)	4.1(5)	4.9(2)
$\log \beta_{021}$	8.4(1)	8.4(1)	18(3)	8.7(2)	13(3)	8.4(2)
$10^4 \cdot U$	528.4	535.4	496.9	634.3	485.1	502.7

· means the correct value

Table 11.

As you can see, the correlation of the estimates of the parameters $\boldsymbol{\beta}$ and $\mathbf{q}$ makes the analysis almost impossible.

¶ 1 calibration function of type -6 is used

† if the calibration function of type -4 were used ($\alpha = 1$, specifier d set to the reciprocal analytical concentration C_3 of Cd^{2+} in each of the 9 solutions $d = 1/C_3 \equiv 1/\mathcal{C}$), the estimate of any of the stoichiometric coefficients of Cd^{2+} would be incorrect ($\mathcal{C}$ is a function of $\mathbf{q}$)

3 References

1. Kývala M., Lukeš I., CHEMOMETRICS '95, Pardubice, Czech Republic, 1995, Book of Abstracts p. 63
2. Havel J., Meloun M., *Talanta* **33** (1986) 525
3. Kankare J. J., *Anal. Chem.* **42** (1970) 1322
4. Schwarz G., *Ann. Statist.* **6** (1978) 461
5. Rissanen J., *Automatica* **14** (1978) 465
3. Hannan E. J., Quinn B. G., *J. Roy. Statist. Soc.* **B41** (1979) 190
7. Anděl J., Perez M. G., Negrao A. I., *Kybernetika* **17** (1981) 514
8. Stünzi H., Anderegg G., *J. Coord. Chem.* **7** (1978) 239
9. Gans P., Sabatini A., Vacca A., *Talanta* **43** (1996) 1739
10. Stypinski B., Dissertation ETH Nr. 11713, Zürich, 1996
11. Anderegg G., personal communication
12. Gampp H., Maeder M., Meyer C. J., Zuberbühler A. D., *Talanta* **32** (1985) 257
13. Havel J., *Pure Appl. Chem.* **34** (1972) 369
14. Frassineti C., Ghelli S., Gans P., Sabatini A., Moruzzi M. S., Vacca A., *Anal. Biochem.* **231** (1995) 374
15. `http://www.chim1.unifi.it/group/vacsab/hypnmr.htm`
16. Ollig J., Hägele G., *Computers Chem.* **19** (1995) 287
17. Hägele G., personal communication
18. Huskens J., Lammers H., Bekkum H. van, Peters J. A., *Magn. Reson. Chem.* **32** (1994) 691
19. `http://www.natur.cuni.cz/~kyvala/opium.html`
20. `http://www.chemi.muni.cz/~lubal`